Environmental Health Perspectives
Vol 90, pp. 181-187, 1991

Daily Diaries of Respiratory Symptoms and Air Pollution: Methodological Issues and Results

by Joel Schwartz,* David Wypij,† Douglas Dockery,† James Ware,† Scott Zeger,‡ John Spengler,† and Benjamin Ferris, Jr.†

Daily diaries of respiratory symptoms are a powerful technique for detecting acute effects of air pollution exposure. While conceptually simple, these diary studies can be difficult to analyze. The daily symptom rates are highly correlated, even after adjustment for covariates, and this lack of independence must be considered in the analysis. Possible approaches include the use of incidence instead of prevalence rates and autoregressive models. Heterogeneity among subjects also induces dependencies in the data. These can be addressed by stratification and by two-stage models such as those developed by Korn and Whittemore. These approaches have been applied to two data sets: a cohort of school children participating in the Harvard Six Cities Study and a cohort of student nurses in Los Angeles. Both data sets provide evidence of autocorrelation and heterogeneity. Controlling for autocorrelation corrects the precision estimates, and because diary data are usually positively autocorrelated, this leads to larger variance estimates. Controlling for heterogeneity among subjects appears to increase the effect sizes for air pollution exposure. Preliminary results indicate associations between sulfur dioxide and cough incidence in children and between nitrogen dioxide and phlegm incidence in student nurses.

Introduction

Quantitative risk assessment for criteria air pollutants raises problems and issues that differ substantially from those involved in assessment of carcinogens. Because there may be thresholds for the effects of air pollutants on the lung, high/dose animal exposure data cannot be easily extrapolated to human exposure in the dose regimes of interest. While animal studies are quite important in identifying possible effects and mechanisms, human epidemiology is central to determining whether there are effects at current ambient concentrations. Epidemiologic studies have the advantage of being in the species and exposure range of interest. They have the disadvantage of introducing greater potential for confounding. The trade-off involves different types of uncertainties. Extrapolation of animal studies over orders of magnitude of exposure and across species introduces great uncertainty into effect estimates. The greater certainty in effect estimates afforded by epidemiology studies of criteria air pollution is countered by an increased uncertainty about whether the effect exists at all, or is due to or hidden by unobserved confounding factors.

Diary studies, defined broadly as studies that record the health status of each study participant repeatedly over time, provide a powerful method for assessing the impact of short-term changes in the environment on human health. If health status is reported as the presence or absence of each of several symptoms, the data consist of sets of sequences of binary outcomes, one for each symptom and participant. The basic analytic objective, to estimate the exposure-response model linking exposure and symptom status, is complicated by the dependencies among responses on successive days (autocorrelation) and among responses of the same subject on different days (heterogeneity). This paper illustrates methods for analyzing diary data that address these complications and demonstrates their use by analyzing data collected in two diary studies, one in children, and the other in nursing students.

The methods described in this report model the incidence rather than the prevalence of symptoms. Incidence is defined as a positive report of symptom occurrence by an individual who did not report that symptom on the previous day. This strategy was chosen because the risk factors for acquiring an illess are not necessarily the same as those that increase its duration. In addition, the use of incidence as an end point greatly reduces, but does not eliminate, the autocorrelation in the data. The low prevalence rates in our data precluded a separate model for the relationship between air pollution and duration.

When the end point is symptom incidence, only subjects free of the symptom on the previous day are at risk. This suggests a relatively simple analysis, in which response rates on successive

*U.S. Environmental Protection Agency, 401 M. Street SW, Washington, DC 20460.
†Harvard School of Public Health, 665 Huntington, Boston, MA 02115.
‡Johns Hopkins School of Hygiene and Public Health, 615 N. Wolfe Street, Baltimore, MD 21205.
Address reprint requests to J. Schwartz, PM221, U.S. Environmental Protection Agency, 401 M. Street SW, Washington, DC 20460.

days were treated as independent observations. This type of analysis will be called ordinary logistic regression in this paper.

Further investigation of the residuals from ordinary logistic regression established, however, that the incidence rates had detectable autocorrelation. Similarly, it is reasonable to expect variability among subjects in the frequency of symptoms and, possibly, in sensitivity to air pollution exposures. This variability may be due to measurable risk factors, such as passive smoking, gas stoves, and the presence of allergic conditions. We refer to such risk factors as subject covariates. Methods for modeling such dependencies are decribed in the modeling section and illustrated in "Results."

Description of the Data Sets

Six Cities Diary Study

The Six Cities Study of Air Pollution and Health is a large longitudinal study of the effects of exposure to air pollutants on the respiratory health of both children and adults (*1,2*). A cohort of approximately 1800 children from six cities (Waterton, Kingston-Harriman, TN; St. Louis, MO; Portage, WI; Stuebenville, OH; and Topeka, KS) was enrolled in a year-long diary study in which parents completed a daily report on the child's respiratory (and other) symptoms. For logistical reasons, the diary study extended over 4 school years (1984–1988). Air pollution concentrations were measured daily in each city during the study. Information on parental smoking, type of cooking stove, and the child's respiratory illness history were obtained via questionnaire.

The diary responses examined were upper respiratory illness (URI) (any two of hoarseness, sore throat, and fever), lower respiratory illness (LRI) (any two of cough, chest pain, phlegm, and wheeze), simple cough (without other symptoms), and any cough (with or without other symptoms).

The incidence rates for all of the symptoms and symptom complexes were low ranging from 0.2% (URI) to 1% (any cough). This implied that the data on recurrence of symptoms on days subsequent to a first report were very sparse. Thus, this report focuses on the analysis of incidence rates. Other more restrictive definitions of incidence rates. Other more restrictive definitions of incidence were considered but had little effect on the analysis because of the consistently low rate of symptom reporting.

Nurses Diary Study

A population beginning nursing school Los Angeles was recruited for a study of viral diseases and other risk factors for acute illness (*3*), Smoking histories and the presence of asthma, hay fever, and other allergic conditions were obtained. Daily diaries of acute respiratory symptoms were handed out and collected each Monday for 3 years. The symptoms examined were headache, cough, sore throat, phlegm, chest discomfort, and eye irritation. Air pollution values were obtained from a monitor within 2.5 miles of the school. Temperature was obtained from a National Oceanic and Atmosphere Administration (NOAA) site within a mile of the monitor. To be eligible for the study, students must be resident at the nursing school. Since they lived, studied, and worked at the same location, there was less mobility than would be found in a general population, leading to more precise exposure estimates. To maintain this exposure profile, subjects were dropped from the study if they moved off campus. Over the course of the 3 years, the size of the study population decreased from over 100 to 35 as students moved away from school.

Models for the Analysis of Diary Data

This section describes methods for analyzing sequences of incidence rates when the objective is to model the effects of temperature, air pollution, and other time-varying variables on the incidence rate. Mismodeling the mean or the covariance structure of the sequences can lead to misleading results about environmental risk.

The data consist of sequences $\{(x_j, Y_j,\ 1 \le j \le T\}$, where $x' = (x_{j1},\ldots,x_{jp})$ is a vector of p covariates affecting all subjects in the study at the *jth* occasion and Y_j is the number of incident cases of the symptom at the *jth* occasion among the n_j subjects who were symptom-free at the previous occasion. Y_j is assumed to have a binomial distribution with parameters n_j and p_j where p_j is the marginal probability of symptom incidence for any subject on day j.

This discussion focuses on the logistic model and its extensions. The logistic model is often used to model binary or binomial outcomes because the parameters can be interpreted as the logarithms of odds ratios and because computing is relatively simple. The logistic model is defined by

$$p_j = \exp[\beta' x_j] / (1 + \exp[\beta' x_j]),$$

or equivalently,

$$\text{logit}(p_j) = \beta' x_j .$$

Both the number of subjects at risk at any occasion, n_j, and the total number of occasions, T, can be large in diary studies.

The goal of the analysis is to estimate the effects of the pollution variables on incidence rates while controlling for other factors, including autocorrelation and subject heterogeneity. Autocorrelation (or serial correlation) refers to the tendency for incidence rates close together in time to be positively correlated. Autocorrelation could be due to state dependence across individuals (e.g., symptoms may occur because other subjects had the symptom on the same or previous days), and/or time-dependent omitted covariates (which tend to be highly correlated in time).

Heterogeneity, or variability among individuals in the probability of response, induces positive correlation among responses on the same individual. Heterogeneity can be due to observable or unobservable within-subject covariates (such as smoking level or illness history), which vary across individuals, or different thresholds, susceptibilities, or reporting behavior across individuals. Differences in reporting behavior could occur, for example, if participants varied in the severity of symptoms considered reportable.

Failure to account for either autocorrelation or heterogeneity in the analysis can lead to errors in inference similar to those resulting from the naive use of standard methods in problems involving misspecified covariates, missing data, or covariate measurement errors. In particular, mismodeling can result in

failure to detect important effects as a consequence of biased point and interval estimates and incorrect hypothesis testing. Diary data typically have positively correlated outcomes, yielding less information than the same number of independent responses, so at a minimum the usual standard error estimates may need to be inflated.

Modeling Autocorrelation

It is natural to begin the analysis of incidence rates with models that assume independence of symptom rates on different days. Preliminary analysis of both diary studies established, however, that residuals from regression models including important covariates were autocorrelate and that this autocorrelation could not be explained by other measured time-varying covariates. Thus, refinements of the model were needed to account for this autocorrelation. This section describes several methods for modeling autocorrelation.

Using Lagged Prevalence or Incidence to Adjust for State Dependence. One possibility is that the probability of symptom occurrence on a given day depends on the participant's symptom status on previous days. When modeling the prevalence of symptoms, Muenz and Rubinstein (*4*), Cox (*5*), and Korn and Whittemore (*6*) used the subject's symptom status on the previous day as a covariate. This approach is not relevant to the analysis of incidence data, however, because all subjects at risk were, by definition, symptom-free on the previous day. As an extension of this idea, however, one could assume that the probability of symptom occurrence for a study participant depends on the symptom status of others in the population on previous days. This dependence could rise if, for example, the symptoms were due to infectious diseases and risk of infection increased with the prevalence of the disease. Such epidemic or clustering effects could be modeled by assuming that

$$p_j = \exp[\beta' x_j + \phi z_j]/(1 + \exp[\beta' x_j + \phi z_j])$$

where z_j, the added covariate, is the lagged prevalence rate in the study population.

The technique of including lagged prevalence rates in the model should be used cautiously, especially when assessing the weak effect of an autocorrelated environmental variable. The pollutant variable under study may also be autocorrelated, and the resulting collinearity will cause bias toward 0 in the coefficient of the pollutant variable if lagged prevalence is added to the model. Adding lagged prevalence or incidence to the model is only justified if there is a biological rationale for doing so, as with certain infectious diseases.

Using Residuals to Modify the Response Probabilities. Observed autocorrelation in incidence rates need not be due to state dependence. Suppose, for example, that there is a time-dependent omitted covriate. In general, such time-dependent variables have an autocorrelation structure of their own that induces autocorrelation in the residuals of the incidence model. As the residuals not only include a random component but are also a function of the omitted variable, the residuals (or a function of the residuals) can serve as a surrogate for the omitted covariate (*7*).

If the n_j are relatively large, using the central limit theorem, we have that approximately

$$Y_j/n_j \sim N(p_j, p_j(1-p_j)/n_j).$$

If the errors $[(Y_j/n_j) - p_j]$ are autocorrelated, modifying the marginal probabilities based on an autoregressive model may be appropriate. As before, let

$$p_j = \exp[\beta' x_j]/(1 + \exp[\beta' x_j])$$

and let

$$p^*_j = p_j + \phi(\sigma_j/\sigma_{j-1})((Y_{j-1}/n_{j-1}) - p_{j-1})$$

where $\sigma_j^2 = p_j(1-p_j)/n_j$. The σ_j/σ_{j-1} term controls for the heteroscedasticity of the symptom rates. Preliminary work suggests that this modification reduces and may eliminate the need for restrictions on the admissable range of values of ϕ, but this issue is still under investigation. These models can be generalized to include second or higher order autoregressive (AR) terms. Because the autoregressive elements are added on the probability scale, we will refer to these models as additive AR models.

Another possibility is to assume an additive contribution on the logit scale. In particular, we could model

$$\text{logit}(p_j) = \beta' x_j + \phi(\sigma_j/\sigma_{j-1})((Y_{j-1}/n_{j-1}) - p_{j-1})$$

This model clearly imposes no restrictions on the allowable range of ϕ. Further obvious modifications could be made to accommodate higher order autoregressive terms. For convenience we refer to these models as multiplicative AR models, since the autoregressive terms occur in the exponent.

In practice, it can be difficult to determine which autoregressive scheme is best. The choice may sometimes be influenced by the statistical software. The choice may influence parameter interpretation. The β parameters have more of a marginal interpretation when the residual effects are added on the probability scale and more of a conditional interpretation if the effects are added on the logit scale. Each of these schemes has the desired effect of reducing autocorrelation of the residuals.

Covariance Models to Accommodate Autorcorrelation Effects. Most, if not all, of the methods described thus far for modeling autocorrelation lead to changes in the interpretation of regression coefficients for the variables under study~ because these coefficients become partial regression coefficients adjusted not only for other covariates but for the residuals included in the model. Obtaining a marginal probability requires integrating over the distribution of the residuals (although this may be trivial for the additive AR model). Liang and Zeger (*8*) and Zeger and Liang (*9*) have described methods for fitting logistic models to the symptom rates while taking account of the correlation among symptom rates on different days. A key fixture of this approach is that the model is marginally logistic, with the autocorrelation in the covariance. This gives their coefficients the usual logistic interpretation. In addition, if multiple time series are available, as in a diary study, the Liang and Zeger approaches yield robust variance estimates that are consistent even if the covariance is

misspecified. The robust estimates are considerably more computationally intensive, however. Liang and Zeger methods can be more efficient than estimators assuming independence of all the observations. The case of first or higher order autocorrelation represents a special case of their method, which promises to be very useful in the analysis of diary data. These robust estimators and covariance matrix estimates deserve more serious attention in many epidemiological applications.

Models for Heterogeneity

This section focuses on the effects of individual heterogeneity, assuming that the residuals are not autocorrelated in time. It is also possible to construct models that combine the autocorrelation and subject heterogeneity effects.

The simplest method for analyzing incidence is to combine the responses from all subjects at each occasion, basically assuming that each subject is identical. This greatly simplifies the computing, which depends only on the order of T, the number of response times, not on the number of subjects. This method may be appropriate when a sample of homogeneous subjects is chosen randomly from a population of interest.

Subject heterogeneity may result from differences in the subject covariates. Since these covariates generally do not change during the diary period, a considerable simplification is obtainable here as well. Collapsing over the T times, we obtain a count W_i, the number of incidents during the period for the *ith* subject. This may be well modeled by a Poisson regression

$$E(W_i) = \exp[\beta' u_i + \ln N_i]$$

where $u_i = [u_{i1}, \ldots, u_{iq}]$ is a vector of q covariates affecting subject i and N_i is the number of days the *ith* subject was eligible to be incident. Again, the computing is considerably simplified, and depends only on the order N of the number of subjects.

If subject covariates are identified that explain differences in individual response rates, the data can be stratified by those covariates. If we assume homogeneity within each strata, which will often be a reasonable assumption, then we can again combine all the responses from all subjects within each strata at each time j. Our model then has responses in the Y_{jk} responses in the *kth* strata at time j out of the n_{jk} subjects at risk. Y_{jk} is assumed to be binomially distributed with parameters n_{jk} and

$$p_{jk} = \exp[\beta' x_j + \tau' v_k]/(1 + \exp[\beta' x_j + \tau' v_k])$$

where x_j denotes the time varying covariates and v_k the stratification variables. Interactions between air pollution and strata are an obvious generalization.

Random Effects Models (Varying Slopes and Intercepts). If T is large, we may observe heterogeneity among subjects in response rates that is not fully explained by the within-subject covariates. Subjects may also vary in their sensitivity to pollutant exposures, as measured by the regression coefficients. Korn and Whittemore (*6*) proposed a two-stage analysis based on the assumption that each subject's sequence of binary responses follows a logistic model but with coefficients that vary among subjects. Specifically, they assume a parameter vector, β_i, for individual i, so that the conditional probability of response for the *ith* subject at the jth response time is given by

$$p_{ij} \mid \beta_i = \exp[\beta_i' x_j]/(1 + \exp[\beta_i' x_j]).$$

They then assume that the β_i arise from a multivariate normal distribution. Their estimation technique also proceeds in two stages. First, estimate $\hat{\beta}_i$ for the *ith* subject using ordinary logistic regression. If the number of observations on the *ith* subject is sufficientiy large, the asymptotic distribution of $\hat{\beta}_i$ is approximated by

$$\hat{\beta}_i \mid \beta_i \sim N(\beta_i, \hat{V}_i),$$

where $\hat{V}_i$ is the usual information-based variance-covariance matrix. Then, in the second stage, we assume

$$\beta_i \mid \beta, \Sigma \sim N(\beta, \Sigma)$$

and so

$$\hat{\beta}_i \mid \beta, \Sigma \sim N(\beta, \Sigma + \hat{V}_i),$$

and β and Σ are estimated from the averages and sums of squares of the $\hat{\beta}_i$ and $\hat{V}_i$ using the method of moments.

Here β and Σ are the population parameters and are viewed as the primary parameters of interest. The method above indirectly accounts for within-subject covariates through the variation in the coefficients. Weighted least-squares regression could also be used to assess the effects of subject characteristics, such as passive smoking or allergy history, on the individual regression coefficients, $\hat{\beta}_i$.

This two-stage estimation method is relatively easy to implement but has two statistical drawbacks, in addition to the computational intensity. First, the asymptotic normality assumption of $\hat{\beta}_i \mid \beta_i \sim N(\beta_i, \hat{V}_i)$ holds only when there is a sufficiently large number of observations per subject and the response rate for each is sufficiently high. In other cases the model is suspect. In particular, they are not appropriate when response rates are very low, as is the case in the Six Cities Diary Study. In fact, for consistency and asymptotic normality of the estimates, we need that both $T \to \infty$ and $N \to \infty$. Second, this estimation method is not the most efficient. More efficient (but more computationally intensive) multivariate random effects models are available (*10*).

Random Intercept Models (Common Slopes). An alternative approach is to assume that the regression coefficients are constant across subjects but that each subject has a different underlying response rate (as measured by the intercept). This formulation allows individual heterogeneity due to observed or unobserved subject covariates, differences in reporting, or other reasons, but information regarding β, the primary parameter vector of interest, is strengthened by combining information across subjects.

In particular, we postulate that responses from the *ith* subject follow the logistic model with success probability

$$p_{ij} \mid \alpha_i = \exp[\alpha_i + \beta' x_j]/(1 + \exp[\alpha_i + \beta' x_j])$$

where α_i denotes the intercept for the *ith* subject.

If one is not interested in the individual intercepts, a conditional maximum likelihood approach can be used. The major

virtue of maximizing the conditional likelihood is that this estimator is consistent and aysmptotially normal in both the large strata and sparse strata cases, i.e., whenever $T \to \infty$ or $N \to \infty$ (or both). The major difficulty is that programs to compute conditional maximum likelihood estimates do not accept the immense amount of data rising from diary studies. For example, if a subject is followed daily for 1 year and has 20 days of symptom incidence, then there are $_{365}C_{20} \sim 4.26 \times 10^{32}$ terms to be summed in the denominator of this subject's contribution to the likelihood.

A second approach is to assume that the individual intercepts, α_i, arise from a common distribution such as the normal distribution and to use a mixture model for the random effects. Here, however, the likelihood will involve integration that cannot be performed analytically, and it becomes computationally intensive to approximate the integrals. Alternative approaches for varying intercept models that involve easier computation and allow estimation of the individual intercepts are under development.

Results

This section illustrates some of the methods discussed in the previous section by applying them to the two sets of diary data. The analyses from the Six Cities Diary are based on data from three cities, Watertown, MA; Kingston-Harriman, TN; and St. Louis, MO. Data for the other three cities are still being collected and processed.

Evidence for Autocorrelation of Incidence Rates

Autocorrelation of time series data should be considered only after controlling for the effects of measured covariates. In the Six Cities diary data, only one independent variable, temperature, had strong and consistent effects on symptom rates. In all analyses discussed in this report, the effects of temperature were controlled by introducing temperature and the square of temperature into the regression model. Each pollutant was investigated separately while controlling for the effects of temperature in this way. Analyses not reported here established that seasonal variables did not contribute significantly to the model after the two temperature terms had been added. Here we consider one set of analyses, those investigating the effects of sulfur dioxide concentration on the incidence of any cough in Watertown.

Figure 1 shows the partial autocorrelation function of the daily incidence rates. The partial autocorrelation of order k *is* the correlation between y_j and y_{j-k} after controlling for $y_{j-1} \ldots y_{j-k+1}$. The magnitude of each bar represents the partial correlation coefficient at that lag. Figure 2 shows the partial autocorrelation function of residuals from an ordinary logistic regression model for cough incidence including temperature, temperature squared, and sulfur dioxide concentration. Autocorrelation is reduced by inclusion of the explanatory variables, but there is a strong indication of, at a minimum, first- and second-order autocorrelation in the residuals. The autocorrelation may be due to unmeasured time-dependent covariates. Epidemic effects, which can be represented as lagged values of prevalence, may also be important. The second panel in Figure 2 shows the partial autocorrelation function after fitting a regression modell with

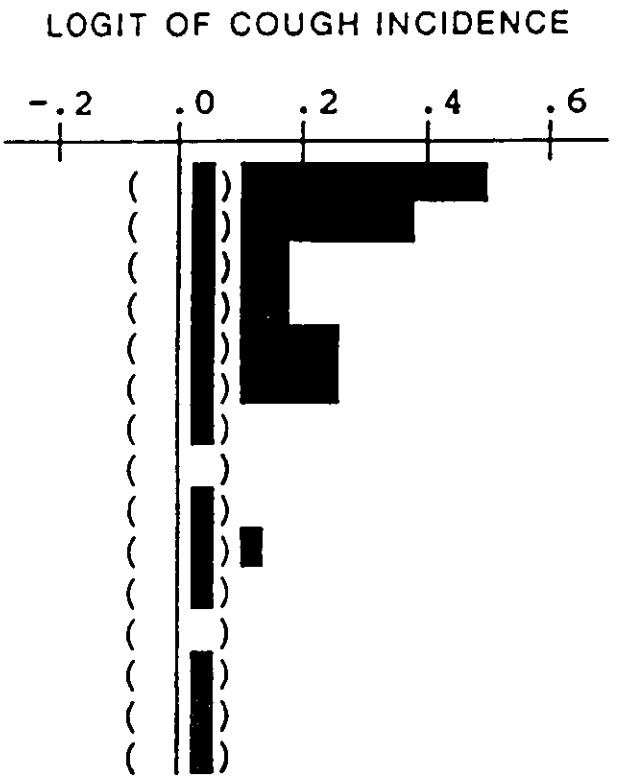

FIGURE 1. Sample correlation function of the daily incidence rates of cough.

first-order autoregressive errors to the data (using the multiplicative AR model). This plot suggests the presence of second order autocorrelation.

As noted in the modeling section, this autocorrelation can be modeled in several ways. Three approaches considered here are the additive AR model, the multiplicative AR model, and the Liang-Zeger model, which were discussed previously. We have also considered the possibility that the symptom probabilities

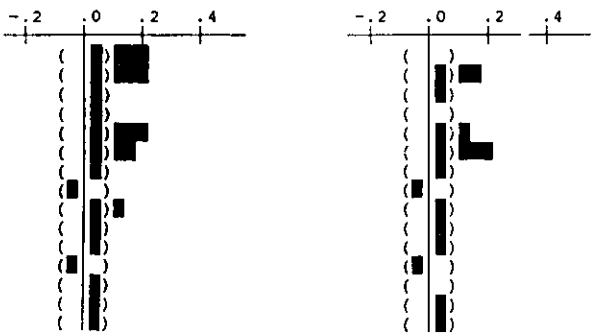

FIGURE 2. The left panel shows the sample partial autocrrelation function of the residuals from an ordinary logistic regression of cough incidence rates on temperature, the square of temperature, and sulfer diozide concentrations. The right panel shows the autocorrelation function when the logistic regression function is modified to assume that the errors have first-order autoregressive error structure.

Table 1. Regression coefficients for cough incidence on sulfur dioxide in Watertown, MA, for different model specifications.

Model	β_{SO_2}	SE	Comments
Ordinary logistic	0.0133	0.0052	
Models with two autoregressive terms			
Multiplicative AR	0.0116	0.0053	
Liang and Zeger (*8*)	0.0132	0.0059	AR(1) insignificant
Additive AR	0.0117	0.0053	AR(1) insignificant
Lagged prevalence [No AR(1) term]	0.010	0.0056	Lagged prevalence insignificant
Reduced models			
Liang and Zeger (*8*) [AR(2) only]	0.0130	0.0059	
Additive AR [AR(2) only]	0.0113	0.0052	

depend on the previous day's disease prevalence.

Table 1 shows the effect of choosing several different models for the error structure on the regression coefficient for sulfur dioxide concentration. Sulfur dioxide was a significant predictor of cough incidence in an ordinary logistic regression model assuming independent errors. Models that adjusted for autoregressive errors tended to reduce the statistical significance of the sulfur dioxide coefficient. In the multiplicative AR model, a first-order autoregressive term was significant but lagged prevalence was not, suggesting that the autoregressive model satisfactorily explains the dependency of the incidence rate on the previous day's outcomes. In the Liang and Zeger and additive AR models, the first-order term was not significant (Table 1). Since these models are slightly diiferent, it is not surprising that they give different results for the order of the autoregression.

Perhaps the most important feature of Table 1 is the consistency among the estimated regression coefficients and standard errors for sulfur dioxide obtained by different methods. Even though the autocorrelation among successive days was of moderate size and highly significant, different approaches to modeling this autocorrelation, including ignoring it entirely (as ordinarily logistic regression does), had little effect on the results.

Individual Effects

One potentially attractive way to account for individual variability is to perform separate regressions on each subject. We call this the Korn and Whittemore (KW) approach. Despite the reservations described in the methods section, we examined this approach for our data. KW also allowed us to examine the relations between individual intercepts and child-specific covariates, such as presence of chronic respiratory disease and parental smoking. In Watertown, the weighted mean of the individual sulfur dioxide coefficients for cough incidence was close to the coefficient obtained from the analysis of the daily incidence rates. In Kingston and St. Louis, the KW approach showed a stronger association than the grouped analysis. This raises the possibility that methods allowing individual variation to weak effects are more sensitive than methods for analyzing pooled data. Nevertheless, the low incidence rates made KW inappropriate for these data. The individual regressions failed to converge for about 10% of the subjects. In the remaining subjects, the distributions of the coefficients were clearly not normal. Even after adjusting for the different weights assigned to different coefficients, a highly skewed distribution remained. Therefore, estimation and testing procedures based on normality assumptions do not apply.

Table 2. Impact of controlling for heterogeneity on the relationship between nitrogen dioxide and phlegm.

Variable*	Basic model β	Stratified model β
Intercept	-2.98	-2.77
Temperature	-0.0124	-0.0170
NO_2	0.756	0.948
Smoking	—	*0.199*

*All variables significant, $p < 0.01$.

For low incidence data, the stratification approach seems better suited. To illustrate this, we use data from the nurses diary study. In Poisson models of subject covariates, smoking, but not prior illness/allergies, was significantly associated with the number of incidents of phlegm each subject suffered. The data were therefore stratified into nonsmokers, subjects with pack-years less than or equal to the median, and subjects with more than the median number of pack-years. Logistic regressions, as described in the section on heterogeneity, were then estimated. Nitrogen dioxide was the only pollutant significantly associated with phlegm. Table 2 compares the results of an simple logistic regression with those of a logistic regression stratified on smoking. Note that here again, controlling for subject heterogeneity increased the estimated effect size for pollution.

Other Temporal Effects

Lag Effects. Any effect of pollution exposure on symptoms is not necessarily contemporaneous. The lag between exposure and symptom may also differ among the pollutants, whose modes of action vary. For instance, Dockery et al. (*11*) reported a lag of 1 to 2 weeks between exposure to high levels of particulates and reductions in lung function. In contrast, Spektor et al. (*12*) and Kinney et al. (*13*) reported that high ozone exposure causes almost immediate reductions in lung function. To explore the lag relationship in the diary data, we used simple logistic regression with no autoregressive components. Temperature was modeled with a linear and quadratic term, as suggested by exploratory plots and analyses. The concurrent and lagged pollutant measures for up to 14 days lag were examined individually. If the pattern in these individual regressions suggested a model using a weighted linear combination of pollutant concentrations on several previous days, such a distributed lag model was also fit. This approach was applied in each of three cities (Watertown, Kingston-Harriman, and St. Louis).

These analyses showed the strongest associations between upper respiratory illness and acid measurements from a continuous sulfuric acid sampler (Table 3 shows the regression coefficients at 0, 1, 2, and 3 day lags). The acid measurements on the two previous days had the largest regression coefficients. A model using a weighted combination of concentrations on the three previous days had the largest coefficient, about twice as large as that for any single day.

Table 3. Coefficients of lagged effects of sulfuric acid concentrations on upper respiratory symptoms.

Location	Lag period 0	1	2	3	Distributed lag
Watertown	0.431	0.683	0.690	0.076	1.28
Kingston-Harriman	0.121	0.461	0.258	0.232	1.16
St. Louis	0.171	0.848	0.524	0.227	2.34

Table 4. Logistic regression for phlegm incidence incorporating autocorrelation and heterogeneity.

Variable	β	SE	p-Value
Intercept	−2.379	0.244	< 0.0001
NO_2	0.843	0.343	0.0140
Temperature	−0.0169	0.0037	< 0.0001
Monday	0.626	0.059	< 0.0001
Smoking	0.207	0.059	< 0.001

Day-of-the-Week Effects. Symptom reporting can be elevated on Mondays and depressed on weekends. This is a particular type of serial correlation that can be modeled by AR terms, but is often better dealt with by day-of-the-week dummy variables. We investigated this issue in the nurses diary, and found a significant elevation in reporting phlegm on Monday. No other day was significant. Table 4 shows a final model combining day-of-the-week effects, stratification by smoking, and a Liang-Zeger approach to modeling the autocorrelation in the covariance. Current exposure, rather than lagged exposure, was the better predictor in this case.

Conclusions

The analyses described in this report have shown that rates of incidence of symptoms among participants in a diary study tend to be autocorrelated, perhaps because of epidemic effects and the effects of omitted covariates on response rates. Moreover, our work and the work of others have shown that subjects have heterogeneous response rates. Analyses of diary data should examine the effects of both autocorrelation and heterogeneity on estimated regression coefficients and their standard errors.

These results do show significant relationships between air pollution and symptom reporting, after incorporation of autocorrelation and heterogeneity. The results indicate that daily diaries can be an important tool for examining the relationship between air pollution and human morbidity.

This work was performed at the U.S. Environmental Protection Agency, the Harvard School of Public Health, and the Johns Hopkins School of Hygiene and Public Health. This work was supported in part by Grants ES01108, ES-07142, and ES-0002 from the National Institute of Environmental Health Sciences, Cooperative Agreement CR-811650 from the Environmental Protection Agency, and Contract RP-1001 from the Electric Power Research Institute. D. Dockery was supported by a Mellon Foundation Faculty Development Award. This paper has not been subjected to EPA peer and policy review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

REFERENCES

1. Ferris, B. G. Jr., Speizer, F. E., Spengler, J. D., Dockery, D., Bishop, Y. M. M., Wolfson, M., and Humble, C. Effects of sulfur oxides and respirable particles on human health: methodology and demography of populations in study. Am. Rev. Respir. Dis. 120: 767–779 (1979).
2. Ware, J. H., Spiro, A., Dockery, D. W., Speizer, F. E., and Ferris, B. G. Jr. Passive smoking, gas cooking, and respiratory health of children living in six cites. Am. Rev. Respir. Dis. 129: 366–374 (1984).
3. Hammer, D. I., Hasselbad, V., Portnoy, B., and Wehrle, P. F. Los Angeles student nurse study. Arch. Environ. Health 28: 255–260 (1974).
4. Muenz, L. R., and Rubenstein, L. V. Markov models for covariate dependence of binary sequences. Biometrics 41: 91–101 (1985).
5. Cox. D. R. The Analysis of Binary Data. Metheun, London, 1970.
6. Korn, E. L., and Whittemore, A. S. Methods for analyzing panel studies of acute health effects of air pollution. Biometrics 35: 795–802 (1972).
7. Box, G. E. P., and Jenkins, G. M. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, CA, 1970.
8. Liang, K. L., and Zeger, S. L. Longitudinal data analysis using generalized linear models. Biometrika 73: 13–22 (1979).
9. Zeger, S. L., and Liang, K. Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42: 121–130 (1986).
10. Stiratelli, R., Laird, N., and Ware, J. H. Random-effects models for serial observations with binary response. Biometrics 40: 961–971 (1984).
11. Dockery, D. W., Ware, J. H., Cook, N. R., Speizer, F. E., Herman, S., and Ferris, B. G., Jr. Change in pulmonary function in children associated with air pollution episodes. J. Air Pollut. Control Assoc. 32: 937–942 (1982).
12. Spektor, D. M., Lippman, M., Lioy, P.J., Thurston, G. D., Citak, K., James, D. J., Bock, N., Speizer, F. E., and Hayes, C. Effects of ambient ozone on respiratory function in active, normal children. Am. Rev. Respir. Dis. 137: 313–310 (1988).
13. Kinney, P. L., Ware, J. H., Spengler, J. D., Dockery, D. W., Speizer, F. E., and Ferris, B. G., Jr. Short-term pulmonary function change in association with ozone levels. Am. Rev. Respir. Dis. 139: 56–61 (1989).